MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2019/2020

TDS3551 – DATA MANAGEMENT

(All Sections /Groups)

13 MARCH 2020
9.00 a.m. - 11.00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question Paper consists of 5 pages with 4 Questions only.

2. Attempt **ALL FOUR** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please print all your answers in the Answer Booklet provided.

QUESTION 1

a) Sharding is a technique to spread read write load across multiple machines. It is also known as horizontal partitioning, utilized in NoSQL databases like MongoDB. Explain in brief how sharding works.

[3 marks]

b) The main idea of the Lambda Architecture is to build Big Data systems as a series of layers. What are these layers and what are they responsible for?

[3 marks]

c) For immutability, a fact-based model is recommended where you deconstruct the data into fundamental units called facts. Why is this so?

[2 marks]

d) Fact-based-model captures facts for an object but do not convey the structure behind the data. What can be used to convey the structure and what are its components?

[2 marks]

Continued...

QUESTION 2

a) The Hadoop Distributed File System (HDFS) is specifically designed for Lambda Architecture data storage. How does HDFS support Lambda Architecture requirements?

[2 marks]

b) Given the following GraphML document, draw its corresponding graphical representation.

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
     http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <graph id="G" edgedefault="directed">
    <node id="n0"/>
    <node id="n1"/>
    <node id="n2"/>
    <node id="n3"/>
    <edge directed="false" source="n0" target="n1"/>
    <edge source="n0" target="n3"/>
    <edge source="n1" target="n3"/>
    <edge directed="false" source="n3" target="n2"/>
</graph>
</graphml>
```

[4 marks]

c) Consider the following dataset and code snippets of the mapper and reducer in Python:

```
...

2012-01-01    09:00  San Jose      Men's Clothing       214.05 Amex
2012-01-01    09:00  Fort Worth    Women's Clothing     153.57 Visa
2012-01-01    09:00  San Diego     Music  66.08  Cash
2012-01-01    09:00  Pittsburgh    Pet Supplies  493.51 Discover
2012-01-01    09:00  Omaha  Children's Clothing 235.63 MasterCard
2012-01-01    09:00  Stockton      Men's Clothing       247.18 MasterCard
2012-01-01    09:00  Austin Cameras       379.6  Visa
...
```

```
#mapper
...
for line in sys.stdin:
      data = line.strip().split("\t")
            date, time, store, item, cost, payment = data
            print store, "\t", cost
...
```

Continued...

```
#reducer
...
salesTotal = 0
oldKey = None

for line in sys.stdin:
      data_mapped = line.strip().split("\t")
      thisKey, thisSale = data_mapped

      if oldKey and oldKey != thisKey:
            print oldKey, "\t", salesTotal
            oldKey = thisKey;
            salesTotal = 0

      oldKey = thisKey
      salesTotal += float(thisSale)
...
```

Modify the mapper and reducer to find the average sales for each item.

[4 marks]

Continued...

QUESTION 3

a) Lambda Architecture places a set of requirements on the serving layer which are batch writable, scalable, and support random reads and fault-tolerance. Explain these requirements.

[4 marks]

b) The normalization/denormalization dilemma commonly found in Relational Database Management Systems (RDBMS) is solved in the Lambda Architecture through the split between the master dataset and the serving layer. Explain why this is so.

[2 marks]

c) Operations at the data access layer (DAL) can be categorized as push or pull types. Describe how these two differ in relation to the external system and serving layer.

[2 marks]

d) A MapReduce mapper written in Python to perform word count is given. Similarly, the same results can also be achieved in Hive. Complete the queries in the given Hive program.

```
#!/usr/bin/python

import sys

for line in sys.stdin:
  line = line.strip()
  words = line.split()
  for word in words:
     print '{0}\t{1}'.format(word, 1)
```

```
CREATE TABLE docs (line STRING);
LOAD DATA INPATH 'input_file' OVERWRITE INTO TABLE docs;
CREATE TABLE word_counts AS

                    :
```

[2 marks]

Continued...

QUESTION 4

a) In the speed layer of the Lambda Architecture, real time views are updated synchronously or asynchronously. Contrast between the two.

[2 marks]

b) Queues are important not only to control flows but also ensure proper operations in asynchronous systems. What is the difference between single-consumer and multi-consumer queues in terms of event disposal?

[2 marks]

c) Two popular models of stream processing are one-at-a-time and micro-batched. Why would you choose one over the other?

[2 marks]

d) The speed layer produces real time views that represent recent data and is continuously updated at a rapid pace. Assuming the system receives 2 GB of data a day, and the data arrives in an average size of 1 MB.

i) If the batch layer generates batch views 4 times a day (every 6 hours), how many times does the speed layer need to perform recomputation in a 6 hour window?
ii) What is the size of each computation?

[2 + 2 = 4 marks]

End of Page.